WHAT IS CLAIMED IS:

1. A method for generating a language component vocabulary VC for a speech recognition system having a language vocabulary V of a plurality of word forms, the method comprising the steps of:

partitioning the language vocabulary V into subsets of word forms based on frequencies of occurrence of the respective word forms; and

in at least one of said subsets, splitting word forms having frequencies less than a threshold to thereby generate word form components.

2. The method of claim 1, wherein the frequencies of the word forms are estimated from a given textual corpus.

3. The method of claim 1, wherein said partitioning step includes the sub-step of numerating the plurality of word forms in the language vocabulary V in descending order based on the frequencies associated with each of the plurality of word forms.

4. The method of claim 1, wherein said partitioning step partitions the language vocabulary V into at least two subsets S1 and S2, and said splitting step splits the word forms of subset S2 into 2-tuple components including stems and endings, but does not split the word forms of subset S1.

5. The method of claim 4, wherein said partitioning step further partitions the language vocabulary V into a third subset S3, with word forms therein being split in said splitting step into 3-tuple components including prefixes, stems and endings.

6. The method of claim 1, wherein said splitting is performed subject to a constraint in which a word that contains a given string of letters is prevented from being split within the string if the string of letters corresponds to one phoneme.

7. The method of claim 1, wherein said splitting is performed using a fixed vocabulary and a fixed list of allowable endings, with each word from the fixed vocabulary being split into at least a stem and an ending that is an element of the fixed set of endings, so as to substantially minimize the total number of all stems that are required to split every word in the fixed vocabulary.

8. The method of claim 7, wherein the fixed set of allowable endings includes an empty ending.

9. The method of claim 1, further comprising generating and storing a word form to corresponding word form components table.

10. The method of claim 9, further comprising the step of labeling each of the word form components stored in said table to distinguish between stems, prefixes and endings.

11. The method of claim 1, further comprising the steps of:

generating a map of said word forms to said word form components, said map further including each of a plurality of non-split words as being associated with itself;

filtering a textual corpus using the map to generate a textual component corpus containing the non-split word forms and the word form components of the map;

accumulating the word form components and the non-split word forms generated by said filtering step in an n-gram language model; and

determining counts of n-tuple sets of word form components and word forms to estimate n-gram probabilities for the n-gram language model.

12. The method of claim 11 wherein said filtering step maps every word in the corpus into a n-tuple word form component.

13. A method for use in speech recognition, comprising the steps of:

splitting an acoustic vocabulary comprising baseforms into baseform components and storing said baseform components; and

performing sound to spelling mapping on said baseform components so as to generate a baseform components to word parts table for use in subsequent decoding of speech.

14. The method of claim 13, wherein said acoustic vocabulary is generated from a textual corpus by applying sound to spelling mapping to said textual corpus, and said method further comprises generating a language model vocabulary from said textual corpus.

15. The method of claim 14, further comprising partitioning said language model vocabulary and splitting said partitioned language vocabulary into vocabulary components.

16. The method of claim 15, wherein said steps of splitting said acoustic vocabulary and splitting said partitioned language vocabulary are performed using the same splitting criteria.

17. The method of claim 13, wherein said splitting comprises splitting baseforms of average size lengths into a first number of components and splitting baseforms of relatively longer lengths into a larger number of components.

18. The method of claim 13 wherein said baseform components are generated independently from language model components.

19. The method of claim 13, further comprising:

performing spelling to sound mapping which includes applying a predetermined set of rules to each word in a word string of a textual corpus, with pronunciations of words being obtained from a word to baseform table; and

baseforms stored in said word to baseform table are collected in said acoustic vocabulary.

20. The method of claim 19, further comprising making entries in said baseform components to word parts table by applying spelling to sound mapping to strings of components, said strings of components being obtained by filtering words of said textual corpus.

21. The method of claim 19, further comprising applying said rules to a language model vocabulary so as to produce new word/baseform pairs in said word to baseform table.

22. The method of claim 19 wherein said sound to spelling mapping is performed via an inversion of said set of rules.

23. The method of claim 22 wherein said sound to spelling mapping produces said baseform components to word parts table by utilizing data from said word to baseforms table, the acoustic vocabulary and the stored baseform components.

24. The method of claim 13, wherein said splitting is performed subject to a constraint in which a word that contains a given string of letters is prevented from being split within the string if the string of letters corresponds to one phoneme.

25. The method of claim 13, wherein said splitting is performed using a sorted and fixed vocabulary and a fixed list of allowable endings including an empty ending, with each word from the fixed vocabulary being split into a stem and an ending that is an element of the fixed set of endings, so as to substantially minimize the total number of all stems that are required to split every word in the fixed vocabulary.

26. A method for decoding a speech utterance using language model components and acoustic components, comprising the steps of:

(a) generating from said utterance a stack of baseform component paths;

(b) concatenating baseform components in a path to generate concatenated baseforms, when the concatenated baseform components correspond to a baseform found in an acoustic vocabulary;

(c) mapping said concatenated baseforms into words;

(d) computing language model (LM) scores associated with said words using a language model, and performing further decoding of said utterance based thereupon.

27. The method of claim 26 wherein said step (d) includes:

mapping said words into a string of sub-words;

computing said LM scores for strings of said sub-words; and

attaching said LM scores to words that produced the corresponding strings of sub-words and performing said further decoding based thereupon.

28. The method of claim 26, wherein said step (a) includes the sub-steps of producing, from said utterance, a set of baseform component strings, and generating said stack of baseform component paths from said strings.

29. The method of claim 26, further comprising the steps of:

(e) mapping the baseform components of said path into word parts, when the concatenated baseform components thereof do not form a baseform found in the acoustic vocabulary;

(f) generating a LM score for an n-tuple of said word parts;

(g) designating a concatenated word form as a valid word, if the LM score for the n-tuple of word parts exceeds a specific threshold, and adding the valid word to a word stack for further decoding.

30. The method of claim 26, further comprising producing an N-best list containing a set of N sentences having the highest likelihood scores, said likelihood scores being a measure of how well candidate sentences match acoustic data.

31. The method of claim 26, wherein said language model comprises a mixture of arithmetic and linguistic language model components.

32. The method of claim 31 wherein said language model is configured to split word numbers into smaller numbers via a modular arithmetic method.

33. The method of claim 31 wherein said words comprise k-tuple word forms, and said method further comprises:

mapping said k-tuple word forms into word numbers;

splitting said word numbers into a t-tuple of integers;

if a probability score for said t-tuples of integers exceeds a predetermined threshold, decoding said t-tuple of integers; else splitting said k-tuple of words into L-tuple linguistic components, computing probability scores therefor, and performing decoding based thereon.

34. The method of claim 31, further comprising splitting said words via linguistic splitting based on morphemes.

35. The method of claim 26, further comprising splitting said words via linguist splitting based on morphemes.

36. The method of claim 26, further comprising splitting said words via linguistic splitting based on any one of spelling, phones and morphemes.

37. The method of claim 33 wherein said splitting said k-tuple of words into L-tuple linguistic components comprises linguistic splitting based on any one of spelling, phones and morphemes.

38. The method of claim 26 wherein said LM scores are computed using a smoothing process for linguistic components, said smoothing process comprising:

verifying whether first and second candidate stems of one of said words has the same set of possible endings by comparing stored ending lists for the respective stems with one another.

39. The method of claim 38 wherein said verifying comprises:

counting a number of times each of the endings in a first said ending list associated with said first stem follows the first stem;

counting the number of times each of the endings in a second said ending list associated with said second stem follows said second stem; and

processing counts resulting from said counting in accordance with a predetermined set of conditions, with probabilities for endings being set if said conditions are satisfied.

40. The method of claim 39 wherein said set of conditions comprises:

said first stem must have high counts for all possible endings that follow it and said second stem must have low counts for at least some endings that follow said second stem;

wherein if said set of conditions is satisfied, then the probabilities for endings following said second stem are set as probabilities for these endings to follow said first stem.

41. The method of claim 39 wherein said set of conditions comprises:

both said first and second stems must belong to a particular class.

42. A method for splitting words in a language vocabulary V in an automatic speech recognition system to provide vocabulary compression, wherein the vocabulary V has a fixed size, the method comprising the steps of:

(a) providing a fixed set of allowable endings, including an empty ending;

(b) providing a fixed set of constraints for splitting words into stems;

(c) initializing a split map of words and the corresponding stems and endings by setting a variable t to a predetermined value, and selecting a first word from the fixed vocabulary;

(d) randomly splitting the first word to generate an ending from the fixed list of allowable endings and a stem;

(e) defining and storing a stem set containing the stem generated at said splitting step (d) and a word set containing the first word;

(f) determining whether t is less than the size of the vocabulary V;

(g) obtaining a new word from the vocabulary V, when t is less than the size of the vocabulary V;

(h) determining possible splits for the new word to generate stems and endings therefrom, using the fixed set of allowable endings and the fixed set of constraints;

(i) determining whether there is a split for the new word that generates a previously stored stem of the stem set;

(j) splitting the current word into the previously stored stem and an ending of the set of allowable endings, when there is a split for the new word that generates the previously stored stem of the stem set;

(k) determining whether another previously stored stem in the stem set can be replaced by a new stem generated at step (h), when there is no split for the current word that generates the previously stored stem of the stem set;

(l) redefining the stem set and the split map to include the new stem generated at step (h) in place of the other previously stored stem, when the other previously stored stem can be replaced by the new stem generated at step (h);

(m) redefining the stem set to include any new stem into which the current word may be split and extending the split map to include the current word by splitting the new word into the new stem, when the other previously stored stem in the stem set cannot be replaced by the new stem generated at step (h); and

(n) incrementing t and returning to step (f) if t is less than the size of the vocabulary V.

43. The method of claim 42, further comprising the step of terminating the method if t is not less than the size of the fixed vocabulary.

44. The method of claim 42, wherein said determining step (k) comprises the step of determining whether other words stored in the word set during previous iterations will remain split after such substitution.

45. The method of claim 42, wherein the vocabulary is sorted such that the words in the language vocabulary V are numerated in descending order based on frequencies associated with each of the words.

46. The method of claim 42, wherein step (j) further comprises the step of extending the split map to the new word.

47. The method of claim 42, wherein step (i) generates all possible splits for the new word.